cancers

MDPI

Article

Using Whole Slide Gray Value Map to Predict HER2 Expression and FISH Status in Breast Cancer

Qian Yao [1,†], Wei Hou [1,†], Kaiyuan Wu [2], Yanhua Bai [1], Mengping Long [1], Xinting Diao [1], Ling Jia [1], Dongfeng Niu [1,*] and Xiang Li [2,*]

[1] Department of Pathology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Beijing 100142, China
[2] PingAn Technology, Beijing 100016, China
* Correspondence: dongfengniu@foxmail.com (D.N.); lixiang453@pingan.com.cn (X.L.); Tel.: +86-15801674868 (D.N.); +86-13810256327 (X.L.)
† These authors contributed equally to this work.

Simple Summary: HER2 expression is important for target therapy in breast cancer patients, however, accurate evaluation of HER2 expression is challenging for pathologists owing to the ambiguities and subjectivities of manual scoring. We proposed a deep learning framework using a Whole Slide gray value map and convolutional neural network model to predict HER2 expression level on immunohistochemistry (IHC) assay and predict HER2 gene status on fluorescence in situ hybridization (FISH) assay. Our results indicated that the proposed model is feasible for predicting HER2 expression and gene amplification and achieved high consistency with the experienced pathologists' assessment. This unique HER2 scoring model did not rely on challenging manual intervention and proved to be a simple and robust tool for pathologists to improve the accuracy of HER2 interpretation and provided a clinical aid to target therapy in breast cancer patients.

Abstract: Accurate detection of HER2 expression through immunohistochemistry (IHC) is of great clinical significance in the treatment of breast cancer. However, manual interpretation of HER2 is challenging, due to the interobserver variability among pathologists. We sought to explore a deep learning method to predict HER2 expression level and gene status based on a Whole Slide Image (WSI) of the HER2 IHC section. When applied to 228 invasive breast carcinoma of no special type (IBC-NST) DAB-stained slides, our GrayMap+ convolutional neural network (CNN) model accurately classified HER2 IHC level with mean accuracy 0.952 $\pm$ 0.029 and predicted HER2 FISH status with mean accuracy 0.921 $\pm$ 0.029. Our result also demonstrated strong consistency in HER2 expression score between our system and experienced pathologists (intraclass correlation coefficient (ICC) = 0.903, Cohen's κ = 0.875). The discordant cases were found to be largely caused by high intra-tumor staining heterogeneity in the HER2 IHC group and low copy number in the HER2 FISH group.

Keywords: breast cancer; HER2; artificial intelligence; deep learning; immunohistochemical (IHC) scoring

Citation: Yao, Q.; Hou, W.; Wu, K.; Bai, Y.; Long, M.; Diao, X.; Jia, L.; Niu, D.; Li, X. Using Whole Slide Gray Value Map to Predict HER2 Expression and FISH Status in Breast Cancer. *Cancers* **2022**, *14*, 6233. https://doi.org/10.3390/cancers14246233

Academic Editors: Hamid Khayyam, Ali Madani, Rahele Kafieh and Ali Hekmatnia

Received: 10 November 2022
Accepted: 14 December 2022
Published: 17 December 2022

1. Introduction

Breast cancer is the most diagnosed cancer that seriously threatens the life and health of women all over the world, with high morbidity and mortality rates of 24.5% and 15.5%, respectively [1]. The HER2 (human epidermal growth factor receptor-2) gene, located at chromosome 17q12–21[2], plays an important role in the development of breast cancer. Fifteen to twenty percent of breast cancer patients are HER2 positive, including HER2 gene amplification and/or overexpression. HER2-positive breast cancer has poor clinical outcomes [2,3], but fortunately, there is a targeted drug-Trastuzumab (Herceptin), which can effectively improve the prognosis [4,5]. HER2 gene amplification assessed by in situ

hybridization (ISH) or protein overexpression assessed by IHC remains the primary predictor of responsiveness to HER2- targeted therapies and a key prognostic biomarker in breast cancer [6]. According to the latest American Society of Clinical Oncology (ASCO)/College of American Pathologists (CAP) guideline [6], all newly diagnosed patients with breast cancer must have a HER2 test performed. In routine clinical practice, the IHC test is first performed. The IHC test gives a score of 0, 1+, 2+, or 3+ that measures the amount of HER2 receptor protein on the surface of cells in a breast cancer tissue sample. The 3+ is the strongest staining, with which the patient must be diagnosed as HER2 positive. 2+ is also known as the equivocal level. Fluorescence in situ hybridization (FISH) must be performed to further decide the HER2 status for patients with IHC 2+ score. Therefore, accurate and efficient HER2 IHC evaluation is important for the diagnosis and treatment of breast cancer patients. In the HER2 IHC test, the HER2-receptor protein is commonly stained with 3,3′-diaminobenzidine (DAB), which has a brown color, meanwhile, hematoxylin staining which has blue color is also applied to visualize the cell nuclei. The stained slide is manually accessed by pathologists under the microscope. Although many countries have implemented national testing guidelines to standardize testing procedures and make results more accurate, the procedure is subjective and semi-quantitative and quite often leads to high inter- and intra-observer variation [7–9]. Therefore, there is an urgent need for an objective and consistent HER2 evaluation system.

Many researchers are devoted to developing computer-aided solutions, semi-automatically or fully automatically, to address the ambiguities and subjectivities of manual scoring. Compared to manual scoring, the computer-aided solution can decrease human error, increase the accuracy of diagnosis, reduce the workload of pathologists, and standardize the scoring systems [10,11]. The pathology whole slide images (WSI) have trillions of pixels, which are too large to process in a single-shot end-to-end way, i.e., processing WSI as a traditional image, even on modern computers. Usually, the fully automatic methods have the following three steps: WSI is first split into small size, i.e., 512×512, image patches; then information of single patch image are extracted; and at last single patch information are summarized to conclude the WSI level result. While the semiautomatic methods need pathologists to manually select regions of interest in the WSI. Masmoudi, et al. [12] presented a method for automated assessment of HER2 IHC staining. They first used a linear classification model on the color information of pixels to discriminate the membrane pixels and nuclei pixels, then watershed algorithm and adaptive ellipse fitting were applied to segment the nuclei and cell membrane. At last, slides were classified into one of the three scoring groups based on features describing the membrane staining intensity and completeness. In contrast to Masmoudi et al. work, HER2CONNECT found the distribution of the area of the connected brown color components (the stained membranes) in the core invasive cancer region had a good correlation with the HER2 expression level, therefore can be used to predict HER2 score. Their method reached 92.3% between the software and the score by the pathologist [13]. Ruifrok et al. [14] proposed a color deconvolution method to deconvolute and quantify the contributions of each staining in the histochemical slide. Motivated by the color convolution method, many researchers were devoted to quantifying the gray level of the HER2 IHC slide. ImmunoMembrane, a web-based application, utilized color deconvolution to separate stained membranes and then designed the IM-score, which is the sum of membrane completeness score and membrane intensity score to classify HER2 scores [15]. Kabakci et al. [16] characterized the cell membrane staining intensity in a comprehensive way using the so call Membrane Intensity Histogram (MIH) method which described the distribution of the staining intensity in different directions.

Deep Learning (DL) models are increasingly being used in various application areas such as computer vision, natural language processing, text or image classification, sentiment analysis, recommender systems, user profiling, etc. [17,18]. Compared to handcraft feature engineering, one of the major advantages of the DL model is the automatic learning feature representation and high representability, which bring the DL model much more versatility when dealing with large datasets and complex problems. Saha et al. [11] developed a cell segmentation model using Trapezoidal LSTM units and HER2 scoring based on the

segmented membranes. However, Saha uses 2048 × 2048 patches, rather than the entire WSI. Qaiser et al. [19] also achieved patch-level HER2 scoring with the help of reinforcement learning. Zhen Chen, et al. [20] proposed a Focal-Aware Module to estimate diagnosis-related regions and a Relevance-enhanced Graph Convolutional Network to summarize information extracted from different levels of the original WSI.

Recently DL models are attracting increasing attention to predicting gene expression status using the WSI image [21–24]. The diagnosis label is usually provided at the WSI level, which cannot be treated as a cluster label of the inputs of the underline model. Therefore, multiple instance learning (MIL) is often implemented to overcome the issue. In this paper, we propose a new artificial intelligence (AI) method to predict HER2 protein expression level and gene status using the WSIs. Instead of using a manual strong label of patch level image or using MIL on the slide-level labeled dataset, we first calculate the unsupervised feature for each patch image, i.e., the gray level, the gray level area fraction, and generate a slide-level feature map using the patch-level feature to represent each patch. In this way, we can reduce the input size of the original slide. Then we build a multi-task deep learning model to predict HER2 protein expression level and gene amplification status simultaneously.

2. Material and Methods

Figure 1 shows the workflow of our study.

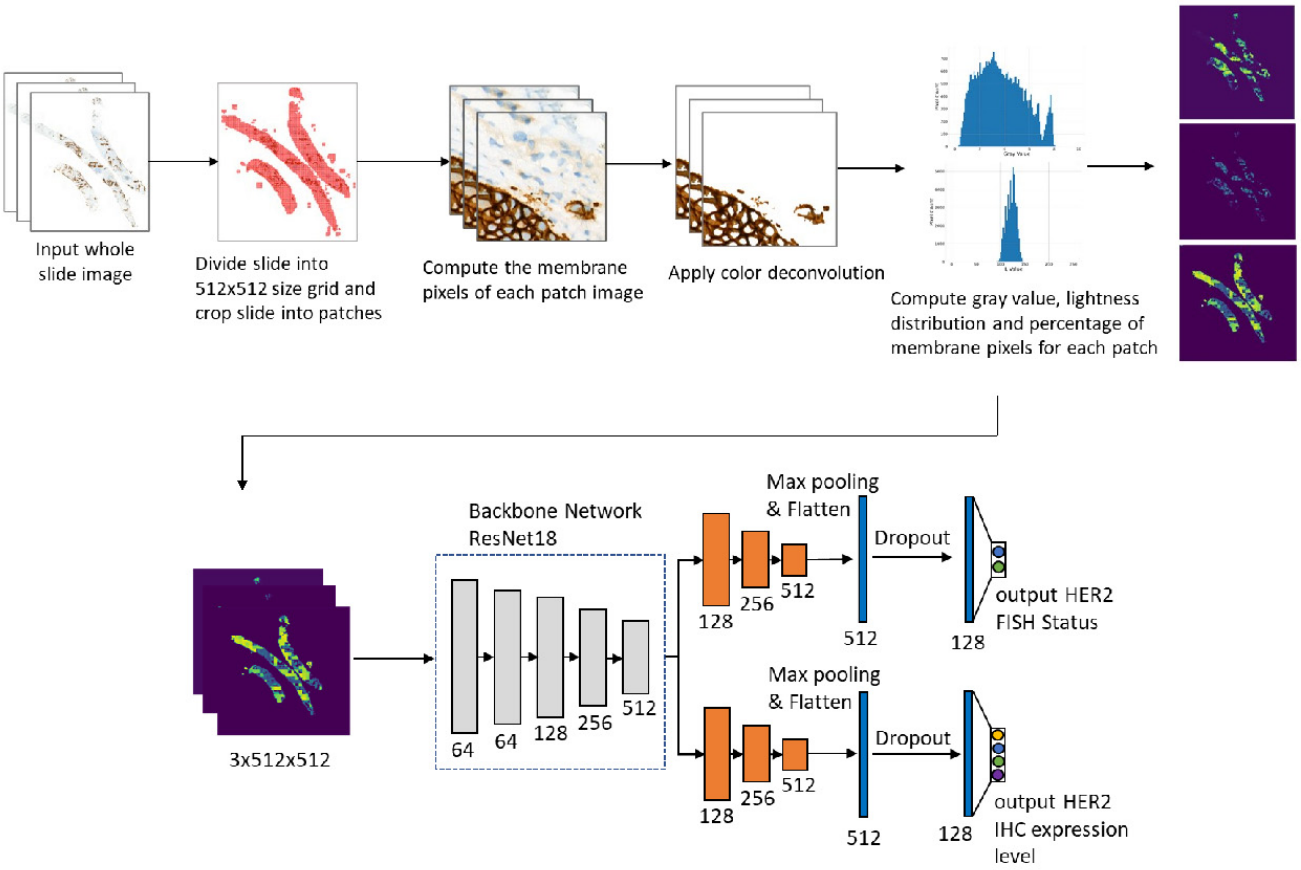

Figure 1. The workflow of our study includes the main steps for preprocessing slides and training the deep learning model. The numbers below the model block give the channel number respectively.

2.1. Human Subjects

We selected 228 biopsy cases of IBC-NST with both IHC and FISH information which were collected between 2010 and 2021 from the department of pathology, Peking University Cancer Hospital & Institute. All subjects were female. Our study obtained permission from

the Peking University Cancer Hospital Institutional Review Board and Ethics Committee (Grant: 2022KT15).

2.2. ImmunohistoChemical Staining

Commercially available primary antibody HER2 (4B5, Roche Ventana) was applied. Immunohistochemical stains were performed on Ventana Benchmark automated immune-Stainer (Tucson, Arizona), following the vendor's protocol. The appropriate positive and negative controls were included for each run. HER2 immunoexpressing was evaluated as 0, 1+, 2+, and 3+ based on the 2018 ASCO/CAP guideline [6] by three experienced pathologists (Q.Y., D.N., and Y.B.). To prevent intra-rater variability, three pathologists were blind to the initial manual evaluation and AI-based scores, and all the cases were reviewed a second time after a 4-week washout period. The discrepant cases were reviewed again to get the final score.

2.3. Fluorescence In Situ Hybridization

HER2 FISH was carried out using the Path Vysion HER2 DNA Probe Kit (Abbott Molecular, Abbott Park, Illinois) and followed the manufacturer's instructions. Two experienced pathologists (DFN and Y.B.) evaluated the HER2 copy number, CEP17 copy number, and their ratios of 20 tumor cells independently and blinded to IHC results. FISH results were recorded as negative and positive according to the 2018 ASCO/CAP guideline. In detail, HER2 FISH results were designated into five groups: group one (G1, HER2/CEP17 ratio $\geq$ 2.0; average HER2 copy number $\geq$ 4.0/cell); group two (G2, HER2/CEP17 ratio $\geq$ 2.0; average HER2 copy number < 4.0/cell); group three (G3, HER2/CEP17 ratio < 2.0; average HER2 copy number $\geq$ 6.0/cell); group four (G4, HER2/CEP17 ratio < 2.0; 4.0 $\leq$ average HER2 copy number < 6.0/cell); and group five (G5, HER2/CEP17 ratio < 2.0; average HER2 copy number < 4.0/cell) [6]. G1 was considered FISH positive and G5 was FISH negative. However, G2 and G4 should evaluate the HER2 IHC results in addition, if not 3+, then those cases should be considered HER2 negative. In G3 cases, when concurrent IHC results are negative (0 or 1+), it is recommended that the specimen be considered HER2 negative.

2.4. Image Processing

The digitized whole-slide images (WSIs) were acquired using a Leica Aperio Versa pathologic scanner (Aperio, Leica Biosystems Imaging, Inc.) viewed at 400$\times$ magnification using Leica ImageScope software. The order of magnitude of pixels was $10^9 \sim 10^{10}$.

Figure 1 shows the flowchart of the method. The whole slide image was first partitioned into 512 $\times$ 512 patches. Then for each small patch image, we segment the membrane pixels using color deconvolution and the k-means method (k-means parameters: number of clusters is 3, the maximum number of iterations is 50, number of redos is 10). After the membrane segmentation, we evaluate the gray value and membrane pixels fraction of each patch. The original WSI is profiled into three maps. In the following, we describe the procedure in detail.

2.5. Membrane Segmentation

The DAB signal is mainly located at the membrane. In the following, we introduce the membrane segmentation method which is based on the color deconvolution and k-means method. Ruifrok etc. applied the Beer-Lambert law to model the stained slide image and proposed the color deconvolution method to separate and quantify immunohistochemical staining [14]. According to the Beer-Lambert law,

$$I_c = I_{0,c} 10^{-AC_c} \quad (1)$$

where I_c is the intensity of light detected after passing the specimen, $I_{0,c}$ is the intensity of light entering the specimen and A is the amount of the stain with absorption factor C. The subscript c indicates the detection channel. By assuming a linear relation between

stain concentration and absorbance, Ruifrok proposed the following color deconvolution method,

$$A = -\log 10\left(\frac{I}{I_0}\right) \times OD^{-1} \tag{2}$$

where A is a vector representing the amount of different stains, I is the transmitted light intensity, i.e., the detected slide image, OD is the normalized optical density matrix, which can be measured experimentally. In the analysis of the HER2 IHC slide, because there are only two kinds of stains, we use the following normalized OD matrix

$$OD = \begin{pmatrix} 0.650 & 0.704 & 0.286 \\ 0.268 & 0.570 & 0.776 \\ 0.636 & -0.710 & 0.302 \end{pmatrix} \tag{3}$$

where the first two row vectors correspond to the OD vectors of hematoxylin and DAB[14] and the last row vector is the normalized cross product of hematoxylin and DAB OD vectors. Following the convention of color deconvolution code given in the Color Deconvolution 2 ImageJ plugin, we use $A = -\log 10\left(\frac{I}{255}\right) \times OD^{-1}$ to deconvolute the original slide image.

After color deconvolution, the value of the 2nd channel corresponds to the intensity of the DAB stain. We then apply the k-means method to the original image. The image is first converted from RGB to Luv color to get better perceptual uniformity which is more suitable for clustering analysis. Define the distance between pixels p, q:

$$D(p,q) = \sqrt{(L_p - L_q)^2 + (u_p - u_q)^2 + (v_p - v_q)^2} \tag{4}$$

where (L_p, u_p, v_p) and (L_q, u_q, v_q) are Luv values of pixel p and q, respectively. Based on the distance $D(p,q)$, we use the k-means algorithm to cluster the pixels in the slice into three clusters, which correspond to the stained cell membrane region, the nuclei region, and the complementary region respectively. At last, we calculate the mean gray values of each pixel group according to the DAB channel calculated previously. We select the group with the highest mean gray value as the cell membrane. Figure 2A–D gives an illustration of the cell membrane segmentation.

2.6. Gray Value Map

In this section, we describe the gray value map which integrates patch-level gray value information to get slide-level gray value information. After segmentation of the cell membrane of each patch image, we calculate the mean gray value and membrane pixel fraction of each patch image. We find that the value of the DAB channel cannot reflect well when the visual gray value is greater than 8, as shown in Figure 2E. By checking the RGB channel value of the membrane pixels, we find that this effect is partially caused by the saturation of the blue channel. It is unclear whether this is truly caused by the stain absorbing all blue light or whether there are some other effects of the hardware device. We notice that the Lightness channel of Luv color space generally reflects the visual gray level except the low gray value range. Therefore, we add the Lightness channel value to the gray value map and build the model to automatically fuse the information. In summary, the gray value A, membrane pixel fraction F, and Lightness value L at patch level are defined as:

$A = \text{mean}_i A_i$ where mean is over all pixels in the membrane cluster,
$F = \frac{\text{number of pixels in membrane cluster}}{\text{total number of pixels}}$,
$L = \text{mean}_i L_i$ where mean is over all pixels in the membrane cluster.
Figure 3 shows the gray value map of IHC HER2 expression 0/1+, 2+, and 3+ cases.

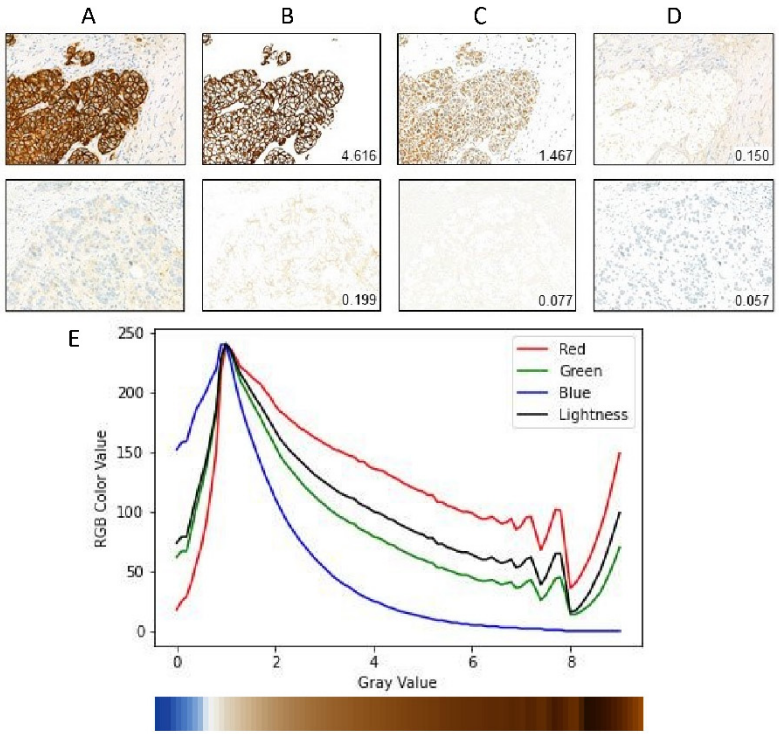

Figure 2. Cell membrane segmentation and the schematic of Graymap. (**A**) raw section of HER2 3+ and HER2 0/1+. (**B**–**D**) are three groups of K-Means output. The gray values are labeled on the images respectively. (**E**) The mean RGB value of different gray value membrane pixels. The bottom color bar is an RGB color map of different gray values.

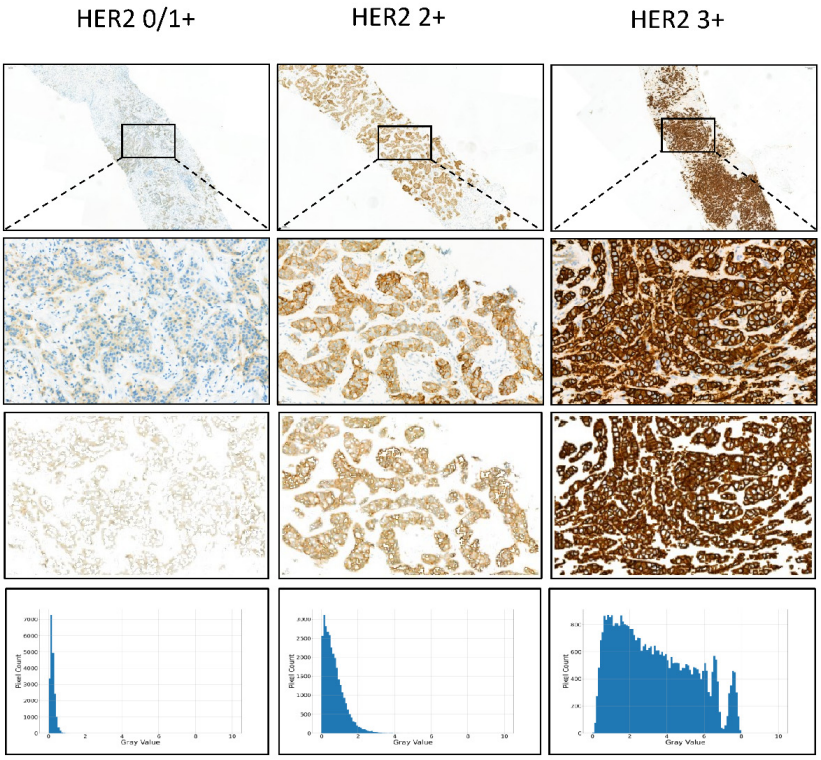

Figure 3. Examples of GrayMap of HER2 IHC expression. Typical examples of HER2 0/1+, 2+, 3+ cases in IBC-NST. From top to bottom: HER2 IHC raw images, magnified images, cell membrane segmentation, and pixels' gray value's distribution of the images.

2.7. Multitask Convolutional Neural Network (CNN)

After getting the gray value map of the whole slide, we further utilize a multi-task CNN model to classify the IHC HER2 expression level and the FISH status simultaneously. We use Resnet18 with base channel number 64 as our backbone network. After the backbone network, we concatenate two task branches corresponding to the IHC HER2 expression classification and the FISH status classification respectively. For each task branch, we use the sigmoid cross-entropy loss as the classification loss and add the dropout layer before the last fully connected layer. All Relu activations are replaced with PRelu to avoid the Relu blow-up issue due to a lack of pretrained weight initialization.

Data augment techniques and manually synthesized images are used to overcome the overfit issue due to the lack of training data samples. We add random rotation (−180, +180), random crop (512, 512) (raw training input size is (680, 680)), random horizontal flip, and random vertical flip data augmentations. We also manually synthesize the image for each original data sample by first manually drawing a mask of a random sample that has the same FISH status, and the same fold-id, but a lower HER2 expression level of the target sample, and then paste the masked part of the selected sample into the target sample's blank space. In this way, we partially increase our training dataset.

The model is implemented in Pytorch using the MMDetection framework and trained with the Adam optimizer with Cosine learning rate policy (learning rate parameters: base learning rate is 0.001, the minimum learning rate is 1.0×10^{-8}). We utilized the 5-fold cross-validation method to evaluate the model. The mean and standard deviation were calculated using prediction on each fold to demonstrate the model performance and stability. Evaluation metrics including precision, recall, F1-score, Jaccard Index, specificity, accuracy, and Area Under Curve of receiver operating characteristic curve (ROC) (AUC) were calculated for binary FISH status prediction. Evaluation metrics including accuracy, F1-score, Cohen's kappa coefficient (κ), and Matthews correlation coefficient (MCC) were calculated for multiclass IHC prediction using macro average mode.

3. Results

3.1. HER2 IHC Status Classification Using GrayMax Model

In the first step, we obtained the manual results of HER2 IHC and HER2 FISH. HER2 IHC was evaluated by three experienced pathologists. We used the median score of three pathologists to further reduce the inter-observer variability, which meant if there was a difference between the three scores, we used the median value of three scores. The details of the HER2 status including IHC and FISH results are shown in Table 1. According to the 2018 ASCO/CAP clinical practice guideline, the cutoff of HER2 IHC staining is 10%, which means the 10% strongest staining of HER2 IHC can be chosen as the represent score of the whole slice. So, we first use the maximum gray value of all patches to represent the gray value of WSI. Then we compared the GrayMax model with the median HER2 scores of pathologists. However, after utilization of the 5-fold cross-validation method, the GrayMax model showed relatively inferior performance with an average accuracy of 0.842 ± 0.023, F1-score of 0.665 ± 0.078, *Cohen's* κ of 0.640 ± 0.063 and MCC of 0.663 ± 0.058 (Table 2). We analysed the details of our model and found the errors in the cases with a heterogeneity of staining, nonspecific cytoplasmic staining, and in cases with invasive micropapillary carcinoma component, mucinous carcinoma component and ductal carcinoma in situ (DCIS) component and interference by necrosis region.

Table 1. Summary of the cohort of the different HER2 statuses.

HER2 Expression Score					
Fish Status	*n*	0	1+	2+	3+
Negative	128	5	19	104	0
Positive	100	0	2	53	45

Table 2. Performance comparison of GrayMax and GrayMap + CNN methods by cross-validation classification.

Method	Fold	Accuracy	F1	Kappa	MCC
GrayMax	0	84.78%	69.71%	67.83%	68.51%
	1	84.78%	70.79%	60.92%	67.47%
	2	86.96%	76.87%	72.23%	73.87%
	3	80.00%	55.38%	53.71%	56.18%
	4	84.44%	59.93%	65.27%	65.49%
	Avg.	84.19%	66.54%	63.99%	66.30%
	Std.	2.28%	7.78%	6.31%	5.77%
GrayMap + CNN	0	93.48%	63.63%	83.13%	84.38%
	1	91.30%	84.65%	80.55%	82.54%
	2	95.65%	94.13%	91.54%	91.96%
	3	100.00%	100.00%	100.00%	100.00%
	4	95.56%	87.81%	90.36%	90.36%
	Avg.	95.20%	86.04%	89.12%	89.85%
	Std.	2.88%	12.39%	6.86%	6.18%

Abbreviation: Avg, Average value; Std, Standard deviation.

3.2. HER2 IHC Status Classification Using GrayMap + CNN Model

To solve the issues of the GrayMax model, we developed a new method to classify the HER2 IHC status. The main issue of the GrayMax model is that a single maximum gray value cannot represent the information of the whole slide. Therefore, we first used the GrayMap of the original whole slide, which contained the gray value information of all the patches, as described in the materials and methods section. Figure 2 showed the segmentation of the cell membrane and the schematic of GrayMap. Figure 3 showed typical examples of GrayMap in a subgroup of 0/1+, 2+, and 3+. Next, we utilized a multi-task CNN model to classify the IHC HER2 expression level as described in the material and methods section (Figure 1). We evaluated the model through a 5-fold cross-validation method and compared the results with three experienced pathologists. The experiment results show that the GrayMap model has much better performance than the GrayMax model with an average accuracy of 0.952 ± 0.029, F1-score of 0.860 ± 0.12, *Cohen's* κ of 0.891 ± 0.069 and MCC of 0.899 ± 0.062 (Table 2). Parameters of evaluation metrics on a subgroup of 0/1+, 2+, and 3+ showed in Figure 4A and Table S1. We further analyzed the intraclass correlation coefficient (ICC) among pathologists and found the ICC value was 0.791 (95% confidence interval [CI], 0.749–0.829) (Figure 4B). It indicated the presence of inter-observer variability and suggested that manual interpretation by the single pathologist may face a high risk of misdiagnosis. Then HER2-AI and HER2-pathologists were compared to show consistency between the AI system and pathologists. The median variables of HER2 pathologists were used in the comparison. The results showed a high consistency between the HER2-AI and HER2-pathologists (ICC = 0.903) (Figure 4C).

3.3. HER2 Gene Status Prediction Using GrayMap+ CNN Model

Since HER2 IHC expression largely represents the HER2 gene amplification status [25]. We also utilized the GrayMap model to predict HER2 gene status and compared the data with the FISH results. Our system demonstrated high performance in predicting HER2 gene status with an accuracy of 0.921, specificity of 0.945, precision of 0.927, recall of 0.89, F1-score of 0.908, and Jaccard Index of 0.832 (Figure 5A and Table S2) and AUC value of 0.936 in the ROC curve which presented the high quality in FISH classification via 5-fold cross-validation method (Figure 5B). This data further confirmed our model as a robust high-performance system not only in HER2 IHC classification but also in HER2 gene status prediction.

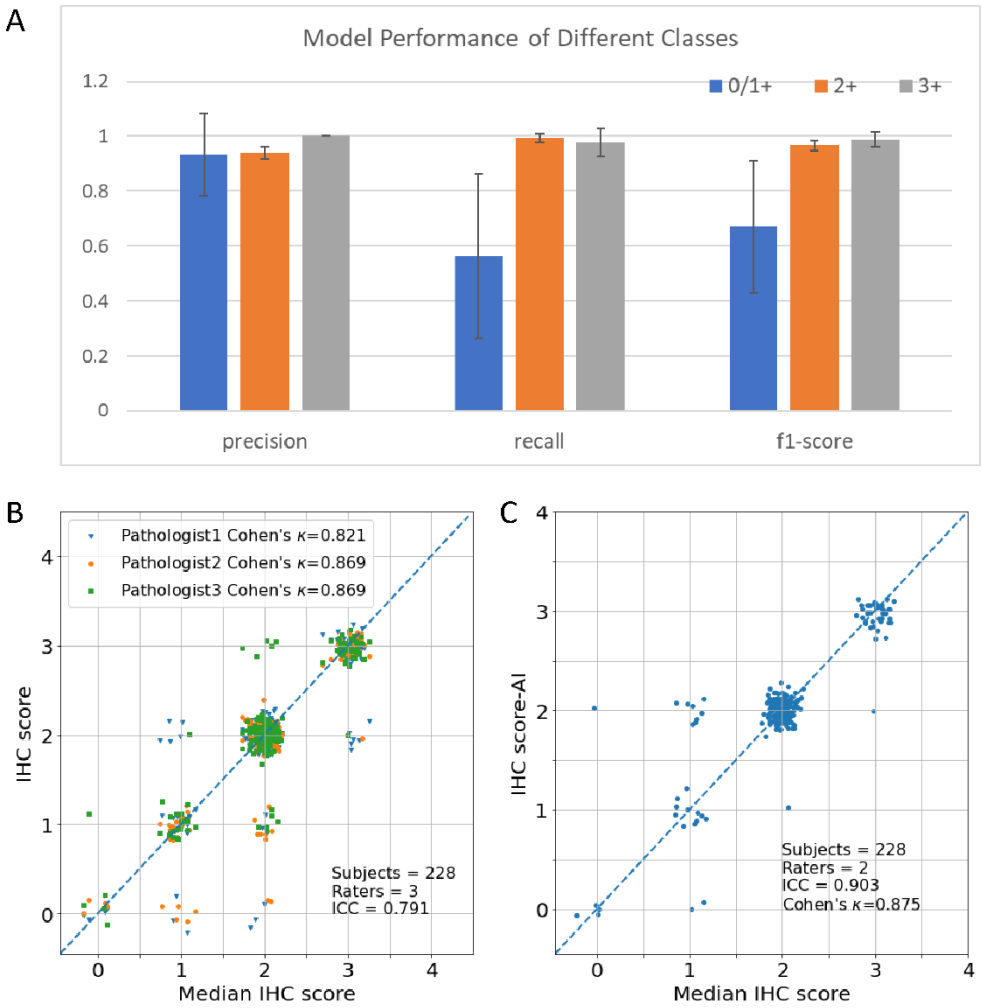

Figure 4. Consistency of the pathologists and the AI system on HER2 IHC classification. (**A**) Histograms of GrayMap model performance in a subgroup of 0/1+, 2+, and 3+. (**B**) The intraclass consistency of HER2 IHC scores in pathologists. (**C**) Consistency of HER2 between AI system (IHC score-AI) and median IHC score in pathologists (median IHC score).

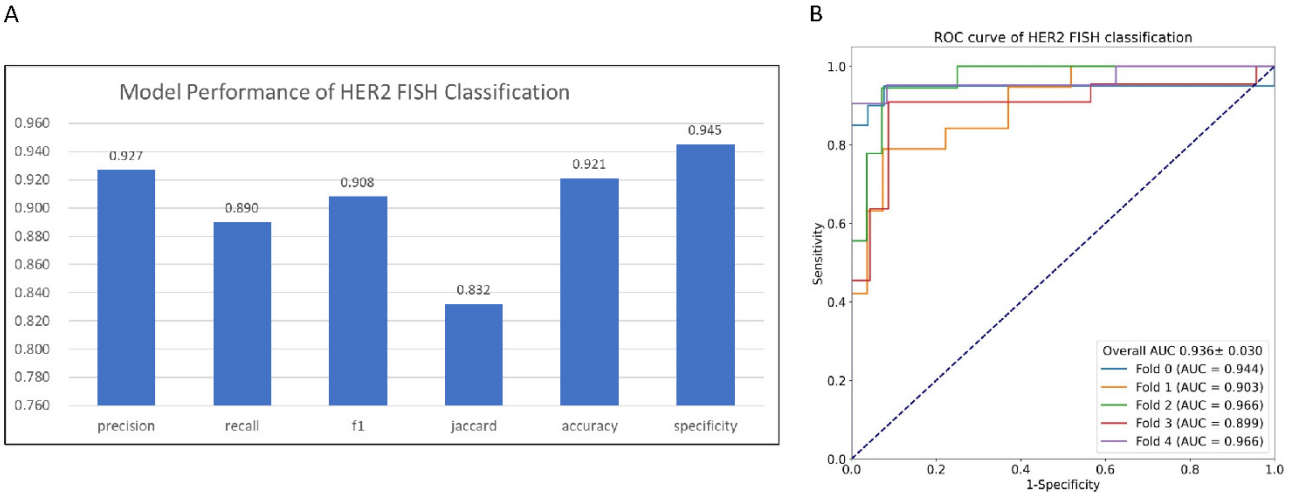

Figure 5. Performance of AI system on HER2 FISH classification. (**A**) Histograms of GrayMap model performance. (**B**) ROC curve of HER2 FISH status classification by cross-validation classification.

3.4. The Analysis of Discordant Cases

The proposed system correctly classified most of the WSIs. However, there were several discordant cases with false positive and negative samples (Figure 6A). We further analyzed the difference between AI systems and pathologists. As for the HER2 IHC results, 13 (13/228, 5.70%) cases were discordant between AI and pathologists. We investigated each case to identify the causes of the variability. Intra-tumor cell heterogeneity of HER2 staining was detected in six cases (6/13, 46.15%) (Figure 6B). Nonspecific cytoplasmic staining was found in four cases (Figure 6C). Another one was due to the nonspecific staining in DCIS (Figure 6D). Our result provided that HER2 staining heterogeneity was identified as the main driver of disagreement between AI and pathologists. Furthermore, the cytoplastic staining can interfere with the machine's extraction of cell membrane staining, resulting in misinterpretation. The nonspecific HER2 expression on DCIS will also lead to error, especially on biopsy tissue with a substantial amount of DCIS. HER2 validation is supposed to be performed only in the IBC-NST component. Since we did not annotate the IBC-NST region on WSIs, we calculated the DCIS component and found 75 cases (75/228, 32.89%) of samples had a DCIS component with a ratio of 5–35%. Only one case (1/75, 1.33%) was included in discordant cases, thus, our model had the ability to resolve the hidden trouble of DCIS. Only two cases could not find a clear explanation for discordance. According to HER2 FISH status, there were 18 (18/228, 7.89%) discordant cases. Five cases were identified intra-tumor cell heterogeneity of dual-color probes. For example, one case with only 2% tumor cells HER2 amplification and one case with 5%. Seven cases have low HER2 copy numbers (average copy number range 4–6 signals/cell). Three cases that were manually evaluated as negative belonged to the G2 and G4 groups, which were the new FISH group according to the 2018 ASCO/CAP guideline. Though the seven low-copy number cases were evaluated as positive and the new FISH group was regarded as negative, the efficacy of HER2-targeted therapy on these groups still needs to be investigated because of the limited evidence with a small subset of cases [6]. Only five cases were left without any explanation for discordance. Our results indicated that AI-based classification guaranteed high diagnostic accuracy and enabled us to reduce misinterpretation.

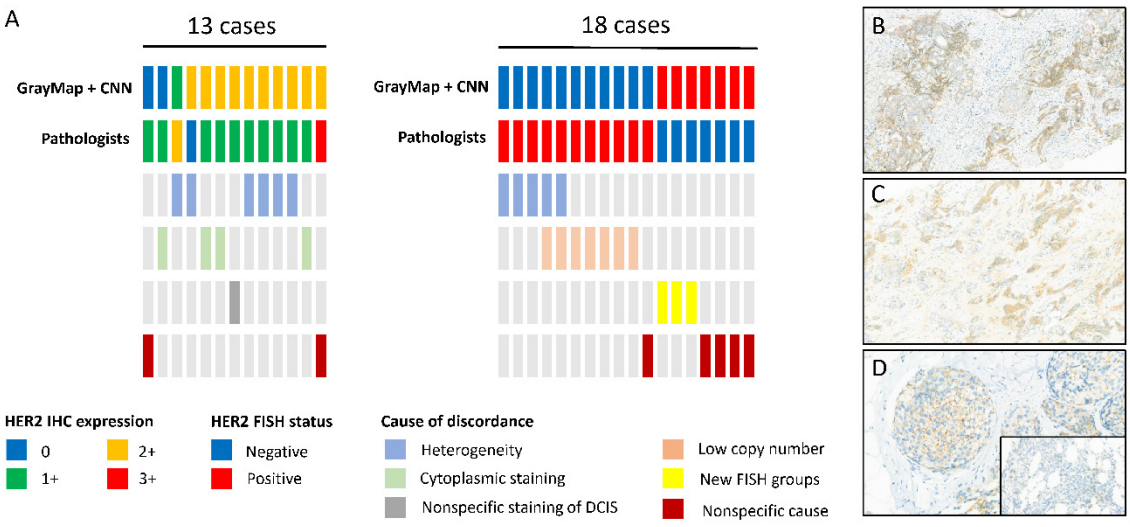

Figure 6. HER2 scoring discordance between pathologists and AI system and the possible causes of the variability. (**A**) Top 2 lines: Comparison between GrayMap model and the pathologist assessment; Bottom 4 lines: The possible causes of the variability; Left: The discordant cases on HER2 IHC classification; Right: The discordant cases on HER2 FISH classification. Vertical bars represent single cases and the representation of different colors are listed at the bottom. The typical image of (**B**) HER2 staining heterogeneity, (**C**) nonspecific cytoplasmic staining, (**D**) nonspecific staining in ductal carcinoma in situ (DCIS) with negative staining of the invasive component.

4. Discussion

In this paper, we proposed a new AI method to tackle the subjectivity and inter-observer disagreement issues of manual interpretation of HER2 IHC slides. The experiments' results showed that the new method could accurately predict HER2 protein expression level (Accuracy 0.95 ± 0.029, Cohen's κ 0.891 ± 0.069) and FISH status (AUC 0.936 ± 0.030). The test of concordance with the three pathologists' interpretation showed that the new method has the highest ICC (ICC 0.903, 95%-Confidence Interval $0.875 \sim 0.924$). Breast cancer (BC) has become the most common cancer diagnosed in women. Personalized medicine, especially drugs focused on target genes in BC, such as trastuzumab, has greatly improved survival. HER2 protein expression level and gene amplification status are the most important indicators for the targeted therapy of BC. However, traditional manual interpretation of HER2 slide has been criticized for subjectivity and inter-observer disagreement among pathologists. This is not only caused by the subjective decision that needs clinic pathologists to take, such as completeness of the membrane staining, intensity of staining, and percentage of positive cells, according to the ASCO/CAP guideline, but also caused by the heterogeneity of BC. AI-based methods, because of the nature of the parametrized model and deterministic behavior, are a prospective approach to solving the pool reproducibility issue of manual interpretation. However, on one hand, the whole slide image is too large to be processed by a single model directly, on the other hand, a single patch-level image of WSI is not able to capture the heterogeneity property of BC. Currently, there are several approaches to solving this issue. The first approach predicts the HER2 expression of each patch and uses the statistical average method to summarize the patch-level results. Compared to this approach, the method proposed in this work adopts a deep learning model to do slide-level predictions, which are more flexible and powerful than the simple statistical average method. Another approach generally follows the ASCO/CAP guideline, making predicting at the cell level. This approach needs considerable human labeling which is not only tedious but also prone to label error, especially for weak staining samples. The weakly Supervised Learning (WSL) method is an attractive method to alleviate patch-level labeling [26]. However, WSL needs a considerable amount of slide-level data. Currently, the performance of WSL on a large HER2 IHC dataset is unclear yet. The method proposed in this work could be another prospective approach to do slide-level predictions.

The proposed AI system can be applied in our actual work in the pathology department. After uploading the WSIs into the system, our model can automatically process patches splitting, cell segmentation, gray value map information extraction, and HER2 IHC and FISH results prediction. The system assists pathologists by pre-reading HER2 IHC slides and presenting calculated results as second opinions to pathologists, especially those with equivocal results as 2+. Our system will significantly mitigate the interobserver discrepancy and contribute to the efficacy and safety of HER2-targeted therapies on BC. At present, a new HER2-low subtype was defined by a score of IHC 1 +or IHC 2+/FISH −, who may benefit from the new HER2-ADC drugs, such as trastuzumab deruxtecan (T-DXd) [27]. The current system has the potential to recognize HER2-low cases with an accurate prediction of both IHC and FISH status.

In our study, compared to the former GrayMax algorithm, the upgraded GrayMap + CNN model can get rid of the most nonspecific and heterogeneous staining problem as well as the special staining pattern of specific breast cancer subtypes in HER2 IHC classification. However, inconsistency between AI systems and pathologists still exists. Consistent with the previous study, HER2 staining heterogeneity was identified as the main driver of disagreement [28]. Intratumoral heterogeneity of HER2 may be due to intrinsic the characteristics of BC, defined as regional heterogeneity and genetic heterogeneity [29]. It may also be caused by IHC procedures, tissue collection, and processing, or slide scanning procedure. In our dataset, most heterogeneity staining cases of the discordant cohort were weak staining thus our model need to improve its capability in dealing with weak HER2 staining. As for HER2 FISH classification, in addition to heterogeneity, a low copy number

(average copy number range 4–6 signals/cell) was the most common cause of inconsistency. According to the 2018 guideline, an average HER2 copy number ≥4 signal/cell is regarded as FISH positive. However, the study showed a clear difference on HER2 copy levels using droplet digital PCR (ddPCR) and targeted next-generation sequencing (NGS) method between the 4–6 copy number groups and ≥6 groups. However, it remains unclear if patients of the 4–6 copy number group derive the same level of benefit as the≥6 groups in HER2-targeted therapy [30]. Futhermore, there were three cases belonging to G2 and G4 groups according to the 2018 guideline, which was the new FISH and should be recognized as FISH negative. However, the researcher showed the G2 group represents a biologically heterogeneous subset, which is different from those in G1 (FISH positive) and G5 (FISH negative) [31]. The G4 group was also proved to be a distinct group with intermediate levels of RNA/protein expression, close to positive/negative cut points [32]. Additional outcome information after HER2-targeted treatment is needed for the new FISH groups.

To improve the accurate, precise, and reproducible interpretation of HER2 IHC results for BC, where quantitative image analysis (QIA) is applied, The College of American Pathologists (CAP) developed the guideline with eleven recommendations [33]. The recommendations suggested that QIA and procedures must be validated before implementation, followed by regular maintenance and ongoing evaluation of quality control and quality assurance. In addition, HER2 QIA performance, interpretation, and reporting should be supervised by pathologists with expertise in QIA. We studied the detailed description of the guideline and found our AI model and procedures met most of the criteria, which suggested the present model is a promising tool for HER2 interpretation. However, this study still had some limitations. First, this work uses the k-means method to segment the cell membrane. It may wrongly classify the cytoplasmic pixels into membrane when the cell is weakly stained or cytoplastic immunohistochemical staining. For most of the weakly stained cases, the method is still able to do correct predictions, because the intensity and percentage of positive cells are major discrimination factors. However, for cytoplastic staining cases, as also demonstrated in the analysis of discordant cases section (four out of 13 total error cases), more local features are needed to discriminate the wrong cases. Secondly, we did not segment the invasive carcinoma region first. The current method relies on the deep learning model to automatically learn features from the data. In future works, we will collect more data and investigate the performance difference between the current method and model which makes predictions only rely on carcinoma region. Third, the completeness of the cell membrane is not represented in the current method. 2018 ASCO/CAP guidelines lay more emphasis on the completeness of cell membrane staining on HER2 2+ and 3+ cases in order to reduce the confusion of pathologists and allow greater discrimination between positive and negative results [6]. Our AI system promised high performance without calculating membrane completeness, however, a feature still needed to be found to represent the completeness of cell membrane staining according to the ASCO/CAP guideline to get a better result.

In conclusion, experimental results indicated that the proposed AI model is feasible for predicting HER2 expression score and HER2 gene amplification using IHC WSI and achieved high consistency with the experienced pathologists' assessments. This unique HER2 scoring model does not rely on challenging manual intervention and is proven to be a simple and robust tool for pathologists to improve the accuracy of HER2 interpretation and provides a clinical aid to target therapy in BC patients.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cancers14246233/s1, Table S1: HER2 IHC classification performance of GrayMap methods by cross-validation classification in the subgroup of 0/1+, 2+, and 3+. Table S2: HER2 FISH prediction performance of GrayMap methods on the subgroup of 0/1+, 2+, and 3+.

Author Contributions: Conceptualization, D.N., K.W., Q.Y., and W.H.; Methodology, X.D., and L.J.; Investigation, W.H., K.W., Y.B., and M.L.; Writing—Original Draft, K.W., Q.Y., and D.N.; Writing—Review & Editing, K.W., and Q.Y.; Funding Acquisition, W.H.; Resources, W.H., and Q.Y.; Supervision, D.N., and X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 81301879 and No. 81702839), Hygiene and Health Development Scientific Research Fostering Plan of Haidian District Beijing (No. HP2022-31-503001), Science Foundation of Peking University Cancer Hospital (No. 2021-11).

Institutional Review Board Statement: This study obtained permission from the Peking University Cancer Hospital Institutional Review Board and Ethics Committee (Grant: 2022KT15).

Informed Consent Statement: The requirements for informed consent were waived due to the noninvasive nature of the study.

Data Availability Statement: All image data associated with this study can be downloaded at https://data.mendeley.com/datasets/3njjk252vc/draft?a=29e5963c-e2d6-4bdb-9c7b-b51d3741b6f0. The source code and the guideline are publicly available at https://github.com/KaiyuanWu/Her2GrayMap. Any further information and requests for resources and materials should be directed to and will be fulfilled by the lead contact, Dongfeng Niu (dongfengniu@foxmail.com).

Conflicts of Interest: The authors declare no conflict of interests.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
2. Slamon, D.J.; Clark, G.M.; Wong, S.G.; Levin, W.J.; Ullrich, A.; McGuire, W.L. Human breast cancer: Correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **1987**, *235*, 177–182. [CrossRef] [PubMed]
3. Tandon, A.K.; Clark, G.M.; Chamness, G.C.; Ullrich, A.; McGuire, W.L. HER-2/neu oncogene protein and prognosis in breast cancer. *J. Clin. Oncol.* **1989**, *7*, 1120–1128. [CrossRef] [PubMed]
4. Cameron, D.; Piccart-Gebhart, M.J.; Gelber, R.D.; Procter, M.; Goldhirsch, A.; de Azambuja, E.; Castro, G., Jr.; Untch, M.; Smith, I.; Gianni, L.; et al. 11 years' follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive early breast cancer: Final analysis of the HERceptin Adjuvant (HERA) trial. *Lancet* **2017**, *389*, 1195–1205. [CrossRef]
5. Woo, J.W.; Lee, K.; Chung, Y.R.; Jang, M.H.; Ahn, S.; Park, S.Y. The updated 2018 American Society of Clinical Oncology/College of American Pathologists guideline on human epidermal growth factor receptor 2 interpretation in breast cancer: Comparison with previous guidelines and clinical significance of the proposed in situ hybridization groups. *Hum. Pathol.* **2020**, *98*, 10–21.
6. Wolff, A.C.; Hammond, M.E.H.; Allison, K.H.; Harvey, B.E.; Mangu, P.B.; Bartlett, J.M.S.; Bilous, M.; Ellis, I.O.; Fitzgibbons, P.; Hanna, W.; et al. Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. *J. Clin. Oncol.* **2018**, *36*, 2105–2122. [CrossRef]
7. Lacroix-Triki, M.; Mathoulin-Pelissier, S.; Ghnassia, J.P.; Macgrogan, G.; Vincent-Salomon, A.; Brouste, V.; Mathieu, M.C.; Roger, P.; Bibeau, F.; Jacquemier, J.; et al. High inter-observer agreement in immunohistochemical evaluation of HER-2/neu expression in breast cancer: A multicentre GEFPICS study. *Eur. J. Cancer* **2006**, *42*, 2946–2953. [CrossRef]
8. Thomson, T.A.; Hayes, M.M.; Spinelli, J.J.; Hilland, E.; Sawrenko, C.; Phillips, D.; Dupuis, B.; Parker, R.L. HER-2/neu in breast cancer: Interobserver variability and performance of immunohistochemistry with 4 antibodies compared with fluorescent in situ hybridization. *Mod. Pathol.* **2001**, *14*, 1079–1086. [CrossRef]
9. Press, M.F.; Sauter, G.; Bernstein, L.; Villalobos, I.E.; Mirlacher, M.; Zhou, J.-Y.; Wardeh, R.; Li, Y.-T.; Guzman, R.; Ma, Y.; et al. Diagnostic evaluation of HER-2 as a molecular target: An assessment of accuracy and reproducibility of laboratory testing in large, prospective, randomized clinical trials. *Clin. Cancer Res.* **2005**, *11*, 6598–6607. [CrossRef]
10. Khameneh, F.D.; Razavi, S.; Kamasak, M. Automated segmentation of cell membranes to evaluate HER2 status in whole slide images using a modified deep learning network. *Comput. Biol. Med.* **2019**, *110*, 164–174. [CrossRef]
11. Saha, M.; Chakraborty, C. Her2Net: A Deep Framework for Semantic Segmentation and Classification of Cell Membranes and Nuclei in Breast Cancer Evaluation. *IEEE Trans. Image Process.* **2018**, *27*, 2189–2200. [CrossRef] [PubMed]
12. Masmoudi, H.; Hewitt, S.M.; Petrick, N.; Myers, K.J.; Gavrielides, M.A. Automated quantitative assessment of HER-2/neu immunohistochemical expression in breast cancer. *IEEE Trans. Med. Imaging* **2009**, *28*, 916–925. [CrossRef] [PubMed]
13. Brügmann, A.; Eld, M.; Lelkaitis, G.; Nielsen, S.; Grunkin, M.; Hansen, J.D.; Foged, N.T.; Vyberg, M. Digital image analysis of membrane connectivity is a robust measure of HER2 immunostains. *Breast Cancer Res. Treat.* **2012**, *132*, 41–49. [CrossRef]

14. Ruifrok, A.C.; Johnston, D.A. Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol.* **2001**, *23*, 291–299.
15. Lodato, R.F.; Maguire, H.C., Jr.; Greene, M.I.; Weiner, D.B.; LiVolsi, V.A. Immunohistochemical evaluation of c-erbB-2 oncogene expression in ductal carcinoma in situ and atypical ductal hyperplasia of the breast. *Mod. Pathol.* **1990**, *3*, 449–454. [PubMed]
16. Kabakci, K.A.; Cakir, A.; Turkmen, I.; Toreyin, B.U.; Capar, A. Automated scoring of CerbB2/HER2 receptors using histogram based analysis of immunohistochemistry breast cancer tissue images. *Biomed Signal Proces. Control* **2021**, *69*, 102924. [CrossRef]
17. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
18. Echle, A.; Rindtorff, N.T.; Brinker, T.J.; Luedde, T.; Pearson, A.T.; Kather, J.N. Deep learning in cancer pathology: A new generation of clinical biomarkers. *Br. J. Cancer* **2021**, *124*, 686–696. [CrossRef]
19. Qaiser, T.; Rajpoot, N.M. Learning Where to See: A Novel Attention Model for Automated Immunohistochemical Scoring. *IEEE Trans. Med. Imaging* **2019**, *38*, 2620–2631. [CrossRef]
20. Chen, Z.; Zhang, J.; Che, S.L.; Huang, J.Z.; Han, X.; Yuan, Y.X. Diagnose Like A Pathologist: Weakly-Supervised Pathologist-Tree Network for Slide-Level Immunohistochemical Scoring. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 47–54. [CrossRef]
21. Chen, M.; Zhang, B.; Topatana, W.; Cao, J.; Zhu, H.; Juengpanich, S.; Mao, Q.; Yu, H.; Cai, X. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *NPJ Precis. Oncol.* **2020**, *4*, 14. [PubMed]
22. Kather, J.N.; Heij, L.R.; Grabsch, H.I.; Loeffler, C.; Echle, A.; Muti, H.S.; Krause, J.; Niehues, J.M.; Sommer, K.A.J.; Bankhead, P.; et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat. Cancer* **2020**, *1*, 789–799. [CrossRef] [PubMed]
23. Wang, X.; Zou, C.; Zhang, Y.; Li, X.; Wang, C.; Ke, F.; Chen, J.; Wang, W.; Wang, D.; Xu, X.; et al. Prediction of BRCA Gene Mutation in Breast Cancer Based on Deep Learning and Histopathology Images. *Front. Genet.* **2021**, *12*, 661109. [CrossRef] [PubMed]
24. Yamashita, R.; Long, J.; Longacre, T.; Peng, L.; Berry, G.; Martin, B.; Higgins, J.; Rubin, D.L.; Shen, J. Deep learning model for the prediction of microsatellite instability in colorectal cancer: A diagnostic study. *Lancet Oncol.* **2021**, *22*, 132–141. [CrossRef]
25. Owens, M.A.; Horten, B.C.; Da Silva, M.M. HER2 amplification ratios by fluorescence in situ hybridization and correlation with immunohistochemistry in a cohort of 6556 breast cancer tissues. *Clin. Breast Cancer* **2004**, *5*, 63–69. [CrossRef]
26. Li, Y.F.; Guo, L.Z.; Zhou, Z.H. Towards Safe Weakly Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 334–346. [CrossRef]
27. Modi, S.; Jacot, W.; Yamashita, T.; Sohn, J.; Vidal, M.; Tokunaga, E.; Tsurutani, J.; Ueno, N.T.; Prat, A.; Chae, Y.S.; et al. Trastuzumab Deruxtecan in Previously Treated HER2-Low Advanced Breast Cancer. *N. Engl. J. Med.* **2022**, *387*, 9–20. [CrossRef]
28. Vandenberghe, M.E.; Scott, M.L.; Scorer, P.W.; Soderberg, M.; Balcerzak, D.; Barker, C. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Sci. Rep.* **2017**, *7*, 45938. [CrossRef]
29. Seol, H.; Lee, H.J.; Choi, Y.; Lee, H.E.; Kim, Y.J.; Kim, J.H.; Kang, E.; Kim, S.-W.; Park, S.Y. Intratumoral heterogeneity of HER2 gene amplification in breast cancer: Its clinicopathological significance. *Mod. Pathol.* **2012**, *25*, 938–948. [CrossRef]
30. Yang, S.R.; Bouhlal, Y.; De La Vega, F.M.; Ballard, M.; Kuo, C.J.; Vilborg, A.; Jensen, G.; Allison, K. Integrated genomic characterization of ERBB2/HER2 alterations in invasive breast carcinoma: A focus on unusual FISH groups. *Mod. Pathol.* **2020**, *33*, 1546–1556. [CrossRef]
31. Wang, X.; Teng, X.; Ding, W.; Sun, K.; Wang, B. A clinicopathological study of 30 breast cancer cases with a HER2/CEP17 ratio of >/=2.0 but an average HER2 copy number of <4.0 signals per cell. *Mod. Pathol.* **2020**, *33*, 1557–1562. [PubMed]
32. Gupta, S.; Neumeister, V.; McGuire, J.; Song, Y.S.; Acs, B.; Ho, K.; Weidler, J.; Wong, W.; Rhees, B.; Bates, M.; et al. Quantitative assessments and clinical outcomes in HER2 equivocal 2018 ASCO/CAP ISH group 4 breast cancer. *NPJ Breast Cancer* **2019**, *5*, 28. [CrossRef] [PubMed]
33. Bui, M.M.; Riben, M.W.; Allison, K.H.; Chlipala, E.; Colasacco, C.; Kahn, A.G.; Lacchetti, C.; Madabhushi, A.; Pantanowitz, L.; Salama, M.E.; et al. Quantitative Image Analysis of Human Epidermal Growth Factor Receptor 2 Immunohistochemistry for Breast Cancer: Guideline From the College of American Pathologists. *Arch. Pathol. Lab. Med.* **2019**, *143*, 1180–1195. [CrossRef] [PubMed]